Fusing Multiseasonal Sentinel-2 Imagery for Urban Land Cover Classification With Multibranch Residual Convolutional Neural Networks

Chunping Qiu, Lichao Mou, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— Exploiting multitemporal Sentinel-2 images for urban land cover classification has become an important research topic, since these images have become globally available at relatively fine temporal resolution, thus offering great potential for large-scale land cover mapping. However, appropriate exploitation of the images needs to address problems such as cloud cover inherent to optical satellite imagery. To this end, we propose a simple yet effective decision-level fusion approach for urban land cover prediction from multiseasonal Sentinel-2 images, using the state-of-the-art residual convolutional neural networks (ResNet). We extensively tested the approach in a cross-validation manner over a seven-city study area in central Europe. Both quantitative and qualitative results demonstrated the superior performance of the proposed fusion approach over several baseline approaches, including observation- and feature-level fusion.**

***Index Terms*— Classification, fusion, long short-term memory (LSTM), multitemporal, nonlocal, residual convolutional neural network (ResNet), Sentinel-2, urban land cover.**

I. Introduction

GLOBAL urban land cover information is a crucial element in various applications, such as efficient infrastructure planning and environmental sanitation improvements, especially in today's rapidly urbanized world, where "55% of the world's population lives in urban areas, a proportion that is expected to increase to 68% by 2050 [1]." Accurate and up-to-date urban information can provide support to decision makers when responding to issues and challenges hampering effective urban governance. In this regard, exemplary existing products include the global urban footprint processed from the TerraSAR-X and TanDEM-X synthetic aperture radar images [2] and the global human settlement built-up grid processed from the Landsat and Sentinel-1 image collections [3]. These remote sensing-based global products give us motivation to explore the potential of satellite images for large-scale land cover mapping. In particular, deep learning has become a powerful technique for such tasks [4].

We have investigated Sentinel-2 images for urban land cover classification in previous studies that were practically aimed at large-scale mapping with openly available data. In addition, the five-day revisit time (at the equator) of Sentinel-2 makes it unprecedentedly feasible for further change detection and long-term monitoring of the urban extent worldwide [5]–[7]. Specifically, we have investigated multiseasonal Sentinel-2 images and demonstrated better results over the single-seasonal input when a residual convolutional neural network (ResNet) architecture was used as the baseline network [8]. Herein, we are motivated to exploit further the potential of fusing the multiseasonal Sentinel-2 imagery for urban land cover classification by proposing a decision-level fusion approach. To provide more methodological insights into this topic, we extensively test the performance of this approach against various baseline fusion approaches.

II. Fusing Multiseasonal Sentinel-2 Images Using ResNet-Based Neural Networks

A. Multibranch ResNet Architecture for Decision-Level Fusion

In this letter, we propose a novel and simple framework to fuse multiseasonal Sentinel-2 images on decision level for urban land cover prediction. The architecture is illustrated in Fig. 1. The architecture mainly consists of a four-stream ResNet to learn the spectral-spatial features from the four seasonal Sentinel-2 images. From the global average pooling of the learned feature maps, urban land cover labels can be predicted independently. Those predictions, i.e., the softmax indicating the class probability, are then averaged and serve as the final prediction. The four-stream ResNet and the averaging part are seamlessly integrated into one architecture without additional inference. Likewise, softmax probability can be predicted from the global average pooling of the feature maps learned by the first three (instead of four) residual blocks of each of the four streams. In this case, all eight predictions can be considered as voters for the final output. Depending on the number of predictions to be averaged (eight or four), the framework is referred accordingly to as Res_ensemble_8 or

Manuscript received June 23, 2019; revised September 16, 2019; accepted October 12, 2019. This work was supported in part by the China Scholarship Council (CSC), in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program under Grant ERC-2016-StG-714087, (So2Sat: Big Data for 4D Global Urban Mapping—10^{16} Bytes from Social Media to EO Satellites), and in part by the Helmholtz Association under the Framework of the Young Investigators Group Signal Processing in Earth Observation (SiPEO) under Grant VH-NG-1018. *(Corresponding author: Xiao Xiang Zhu.)*

Chunping Qiu, Lichao Mou, and Michael Schmitt are with the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: chunping.qiu@tum.de; lichao.mou@dlr.de; m.schmitt@tum.de).

Xiao Xiang Zhu is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2019.2953497

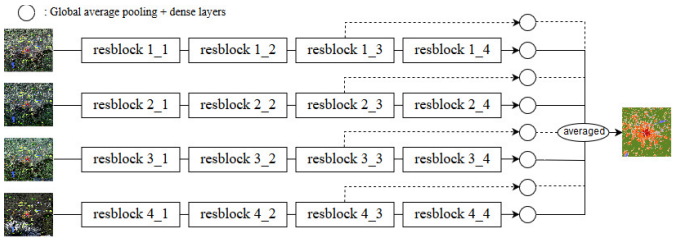


Fig. 1. Fusing network architecture for urban land cover classification. Here, T_1, T_2, T_3, and T_4 indicate four seasons. Weights of different streams are not shared and dropout is not shown. The network is referred to as Res_ensemble_8 and Res_ensemble_4, when predictions from low-level features (the dotted lines) are considered and not considered, respectively. Input is the multiseasonal images and the output is a fused prediction.

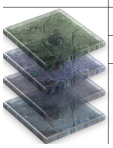
seasonal input	before backbone	backbone	after backbone	baseline appraoches
	stacked as input	ResNet	-	st-ResNet
	non-local block			NL-ResNet
	independent inputs for respective streams	multi-stream ResNet	LSTM	Re-ResNet
			non-local block	ResNet-NL
			averaged softmax	Res_ensemble

Fig. 2. Various approaches to fuse multiseasonal Sentinel-2 images for land cover prediction. The backbone is used to extract features, while modules before and after the backbone are designed to exploit the multiseasonal images.

Res_ensemble_4 in this letter. This way, the different decisions made independently of different seasons can be harnessed for urban land cover classification.

B. Baseline Approaches for Fusing Multiseasonal Sentinel-2 Images

Multiseasonal Sentinel-2 images can be fused on observation- and feature-levels, in addition to the decision-level approach [9].

1) *Observation-Level Fusion:* To be used as the input to a conventional ResNet, the multiseasonal images (along the spectral band dimension) are stacked (i.e., multiseasonal images together are treated simply as one image that contains the ground status of different seasons). This is referred to as st-ResNet in Fig. 2.
2) *Feature-Level Fusion:* Multibranch ResNet is followed by a long short-term memory (LSTM) network, with multiseasonal images as inputs to the independent branches (Re-ResNet in Fig. 2). This way, multitemporal information can be exploited through the extraction of temporal features, which can be the complementary information to the spatial-spectral information learned by the multibranch ResNet. Our previous work has shown that Re-ResNet (feature-level fusion) provides higher accuracy than st-ResNet (observation-level fusion) for mapping urban land cover [8].

Under the above-mentioned multibranch framework, instead of LSTM, the multitemporal information can be alternatively exploited by nonlocal neural networks that have been shown effective for video classification by capturing long-range dependences instead of processing one local neighborhood at a time [10]. We were inspired by its ability to directly capture the spatiotemporal dependences, which might be beneficial to the successful fusion of multiseasonal Sentinel-2 images. Depending on where the nonlocal block was plugged into the network architectures, before or after the ResNet backbone, two types of nonlocal-based architectures, namely, NL-ResNet and ResNet-NL, were explored in this letter.

The overall structures of different fusion approaches, i.e., st-ResNet, NL-ResNet, Re-ResNet, ResNet-NL, and Res_ensemble, are presented in Fig. 2. These approaches were compared and analyzed in detail herein. Note that for each of these approaches, different subnetworks were seamlessly integrated into one architecture so that no postprocessing was needed. Among these approaches, st-ResNet and NL-ResNet are similar in that they both directly model multitemporal dependences using the input multiseasonal Sentinel-2 images, instead of using the learned feature maps. Re-ResNet, ResNet-NL, and Res_ensemble are similar in that they all model multitemporal dependences from the learned features by a multibranch structure (multiple ResNets in parallel).

To draw a valid conclusion, we further investigated the effect of the network depth on the classification accuracy. A shallow version and a deep version of ResNet, with a depth of 14 and 56, respectively, were chosen as the backbones. Their detailed architectures can be found in [8] and [11], respectively.

III. Experimental Results

A. Study Areas and Data Sets

We followed the same experimental setup as in our previous work [8], so as to ensure that our comparison of the different fusion approaches was done in a meaningful manner. Study areas, data sets, and experimental setups were briefly described below, to facilitate complete understanding. The study area comprises seven cities across Europe, namely, Amsterdam, Berlin, Cologne, London, Milan, Munich, and Paris. For each approach, seven experiments were carried out, in each of which one hold-out city is used for test and the other six cities were used for training. This is to test the generalization ability of the models of different approaches. In each city, we processed mostly cloud-free multiseasonal Sentinel-2 images for each of the four seasons from winter 2016/2017 to autumn 2017 using the Google Earth Engine (GEE) [12]. We used the 10-m bands consisting of B2 (blue), B3 (green), B4 (red), and B8 (near-infrared) and the 20-m bands consisting of B5 (red edge 1), B6 (red edge 2), B7 (red edge 3), B8a (red edge 4), B11 (short-wavelength infrared 1), and B12 (short-wavelength infrared 2), upsampled to 10 m. Reference class samples used as ground truth were from the LCZ42 data set [13] and were further prepared by class combination and data augmentation as in [8], to overcome the class-imbalance problem. In addition, accuracy assessment was carried out on absolutely balanced samples. Therefore, only two measures, overall accuracy (OA) and Kappa coefficient, were used for accuracy comparison and analysis.

B. Comparison of Accuracy Resulting From Different Fusion Approaches

Table I provides a list of classification accuracies achieved by each of the different approaches. A comparative evaluation of these accuracies showed these general findings.

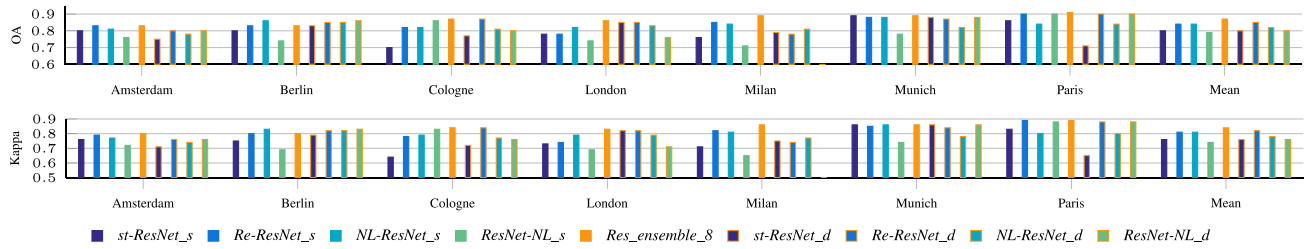


Fig. 3. OA and Kappa coefficient values of seven test cases resulting from nine different fusion approaches. All samples of the test areas are unseen by the respective trained networks and the last column (mean) is the averaged results over all seven test cases corresponding to values in Table I.

1) Decision-level fusion provides the most and second-most accurate classification results with the highest OA and Kappa. Under the decision-level fusion category, a further improvement can be achieved by jointly considering the predictions from both low- and high-level features, i.e., using eight voters instead of four.
2) Re-ResNet_d achieves the most accurate classification results under the feature-level fusion category. In addition, it provides a slight improvement over its shallow version, Re-ResNet.
3) NL-ResNet provides more accurate results than the ResNet-NL, which is true for both their shallow and deep versions.
4) For all three feature-level and one observation-level fusion approaches, deeper networks with more trainable parameters do not necessarily provide considerable benefits. For NL-ResNet, the deeper version provides even worse results.
5) Observation-level fusion provides no benefits with slightly lower OA and Kappa compared with av-ResNet, which does not exploit multitemporal information.

Fig. 3 further provides a detailed comparison of these fusion approaches for all seven test cases. Note the consistency of the above findings in most test cases.

C. Visual Comparison of Classification Maps Resulting From Different Fusion Approaches

We also carried out qualitative comparisons to complement the quantitative results in providing more insights into the characteristics of the different fusion approaches. In particular, we chose two subsets from the Munich, Germany, area as samples, for which urban land cover maps resulting from five representative approaches are presented and compared in Fig. 4. With the manually labeled polygons as reference, Res_ensemble apparently provided more accurate classification results than either NL-ResNet_d or Re-ResNet_d, which classified some open built-up areas as compact built-up areas, as indicated by the noticeable larger red areas. This finding was consistent with that from Fig. 5.

IV. Discussion

In general, the experimental results summarized in Section III already provided a clear answer to the question that motivates this investigation and it gets apparent that decision-level fusion is better for our task given the input data sets and experimental setups. In addition, we achieved promising classification results for seven distinct unseen test areas, which indicate a strong generalization ability of the trained networks. Beyond that, the following major insights can be gained based on the interpretation of the presented experimental results, which can be beneficial to similar tasks at a similar scale.

TABLE I

Comparison of Classification Accuracy Achieved by Different Fusion Approaches. (Measures Are Averaged Over All Seven Test Cases. Bold Values Indicate the Best Accuracy Achieved for the Respective Fusion Levels. The Symbol "_d" Denotes the Deep Versions of Respective Networks)

Fusion	Approach	OA	Kappa	Param.(M)
Not considered	Spring	82.7%	0.79	0.282
	Summer	81.2%	0.77	
	Autumn	**82.7%**	**0.79**	
	Winter	77.9%	0.74	
	av-ResNet	81.1%	0.77	
Observation-level	st-ResNet	79.7%	0.76	0.286
	st-ResNet_d	**79.8%**	0.76	1.668
Feature-level	Re-ResNet	84.1%	0.81	0.844
	NL-ResNet	83.9%	0.81	0.308
	ResNet-NL	78.5%	0.74	1.369
	Re-ResNet_d	**84.6%**	**0.82**	7.175
	NL-ResNet_d	81.8%	0.78	1.690
	ResNet-NL_d	79.6%	0.76	7.179
Decison-level	Res_ensemble_4	85.3%	0.82	1.127
	Res_ensemble_8	**86.7%**	**0.84**	1.163

A. Decision-Level Fusion as the Better Approach

Both the quantitative comparisons measure and the resulting land cover maps in Section III demonstrated the superior performance of decision level over both the observation- and feature-level fusion approaches. With all test cases considered, OA and Kappa can be improved from 84.6% to 86.7% and 0.82 to 0.84, respectively, as compared with the sophisticated Re-ResNet that integrates a four-stream ResNet and an LSTM and that has much more trainable parameters, as shown in Table I. We therefore suggest that decision-level fusion should be considered with high priority over sophisticatedly designed architectures when fusing multiseasonal Sentinel-2 images when the study is application-oriented. A similar finding, i.e., decision-level fusion provides the best result in the context of deep learning, was recently documented in [14] for the fusion of heterogeneous input data (exemplified by the aerial and street-view images) for building-type classification.

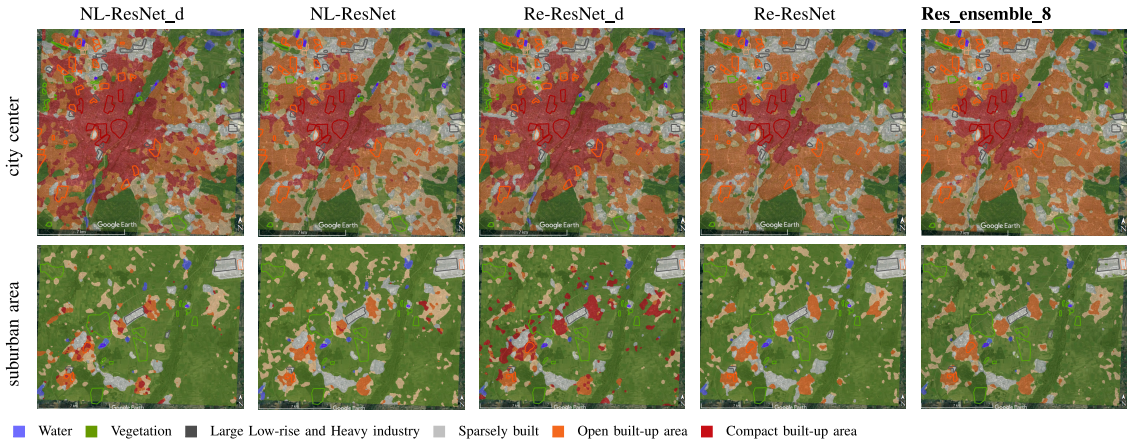


Fig. 4. Comparison of land cover classification maps overlaid on Google Earth images resulting from different approaches, with the city center and suburban area of Munich, Germany, as examples. The polygons are manually labeled for reference with the same legend as the land cover classes. The satellite image data are from Google, Image Landsat/Copernicus.

What has to be mentioned, however, is that in order to address the cloud problem, we have aggregated all images available for each meteorological season into a single, mostly cloud-free image. The preprocessing is already equivalent to an observation-level fusion of all the available Sentinel-2 images within that season. Consequently, the temporal resolution was not fully preserved in the processed Sentinel-2 images. Furthermore, the aimed land cover classes over urban areas herein did not clearly exhibit different phenological stages as crops in the task of crop identification [15]. On the contrary, a fusion on the decision level was able to exploit the joint power of the multiseasonal images robustly.

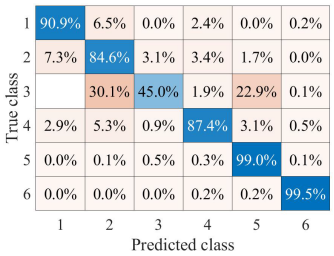


Fig. 5. Combined confusion matrices of the seven test cases resulting from Res_ensemble_8. Confusion matrix was created considering all test samples from all seven test cases together. Classes 1–5 are compact built-up area, open built-up area, sparsely built, large low-rise and heavy industry, vegetation, and water, respectively.

B. Influence of Network Depth

Our experimental results did not confirm the generally correct and well-known rule of "deeper is better" in deep learning. With the same type of architecture, a deeper Re-ResNet with six times more trainable parameters would only result in a slightly higher accuracy, as shown in Table I. In the case of NL-ResNet, a deeper version even would lead to even worse results. The possible explanation for this is that the shallow versions of the used networks are already deep enough to capture the characteristics of our training data, i.e., the size of the training data is not big enough, or the spatial resolution is not high enough to exploit the power of deeper networks fully. Since the achieved test accuracy calculated on completely unseen data is, however, already quite promising, we assume the generalization capability of the trained model to be good enough for large-scale production purposes.

C. Effect of the Nonlocal Block

The nonlocal block-based NL-ResNet was able to provide encouraging classification results that were close to the best achievable results on a feature-level fusion, as shown in both Table I and Fig. 4. In particular, NL-ResNet brought a distinct advantage over st-ResNet without necessarily introducing more parameters, thanks to its inherent ability to capture long-range dependences over both spatial and temporal dimensions. Moreover, note that the nonlocal block was the only difference between the architectures of st-ResNet and NL-ResNet, thus confirming our hypotheses that the nonlocal block is suitable for exploiting the multitemporal information within the Sentinel-2 images for land cover classification. Furthermore, NL-ResNet provided more accurate results than ResNet-NL, which means that its advantages can be gained when using a nonlocal block directly on the original input images than on the learned feature maps in which temporal information might be destroyed.

D. Confusions

In spite of the overall good results, there are still some confusions among the classes remaining, as shown in Fig. 5. Specifically, *sparsely built* was classified as *open built-up area* and *vegetation*. This is understandable, because these classes appear similarly in the Sentinel-2 images. Intuitively, both *open built-up* and *sparsely built area* include buildings and vegetation. In addition, the used patchwise classification approaches tended to be affected by the contextual features learned from the neighboring areas. To improve the classification accuracy over Res_ensemble further, we recommend the following directions: class-specific features can be learned by attention-based neural networks to better distinguish the classes. In addition, differentiation of the different classes and detection of small built-up areas can be enhanced by jointly considering the related tasks in a multitask manner.

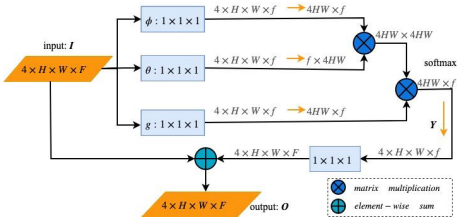


Fig. 6. Nonlocal module for multitemporal information modeling. $1 \times 1 \times 1$ is $1 \times 1 \times 1$ convolution.

V. Conclusion and Outlook

Urban land cover maps can be used as a first step for hierarchical LCZ classification or human settlement extent mapping. Based on our previous studies, we proposed in this letter a new decision-level approach that is capable of fusing multiseasonal Sentinel-2 images, as demonstrated through a set of cross-validations over a seven-city study area in central Europe. Based on a comprehensive comparison with several baseline approaches, we conclude that decision-level fusion is superior over feature-level fusion for similar tasks at a similar scale, when multiseasonal Sentinel-2 images are exploited. We expect the findings of this letter to provide additional insights and pave the way for the realization of large-scale applications, such as land cover and land use classification.

Appendix
Nonlocal Neural Networks

The nonlocal block used in this letter is illustrated in Fig. 6. Given the input feature maps (or image patches) I, the output of a nonlocal block is a sum of the nonlocal-based estimations and the original feature maps to keep the initial behavior of the employed networks

$$O = WY + I \tag{1}$$

in which W is a learnable weight matrix implemented as a $1 \times 1 \times 1$ convolution to compute a positionwise embedding on Y, which is a neural network-based nonlocal operation

$$Y = \sigma\left(I^T W_\theta^T W_\phi I\right) g(I) \tag{2}$$

where W_θ, W_ϕ, and g are all the weight matrices that are to be learned through the implementation of $1 \times 1 \times 1$ convolutions. $W_\phi I$ and $W_\theta I$ transferred the original feature maps into an embedding space, in which patch similarity is modeled for all positions. Such manner of computing similarity is called embedded Gaussian, which is chosen for this letter. Other options include Gaussian, Dot product, and concatenation. The symbol σ in $\sigma(I^T W_\theta^T W_\phi I)$ results from a combination of the normalization factor and computed similarities. In summary, the nonlocal block used in this letter corresponds to a generic nonlocal operation in the following way: $\sigma(I^T W_\theta^T W_\phi I)$ is the weights representing the similarity between each two of all positions along both the spatial and time-step dimensions, and $g(I) = W_g I$ is a linear embedding of the original feature maps.

References

[1] *2018 Revision of World Urbanization Prospects*, United Nations, New York, NY, USA, 2018.
[2] T. Esch *et al.*, "Urban footprint processor—Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1617–1621, Nov. 2013.
[3] M. Pesaresi *et al.*, "Operating procedure for the production of the global human settlement layer from Landsat data of the epochs 1975, 1990, 2000, and 2014," Publications Office Eur. Union, Joint Res. Centre Eur. Commission, Brussels, Belgium, Tech. Rep. EUR 27741 EN, 2016.
[4] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
[5] M. Drusch *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.
[6] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.
[7] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, p. 980, 2018.
[8] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "LCZ-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019.
[9] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, Dec. 2016.
[10] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE CVPR*, Apr. 2018, pp. 7794–7803.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. IEEE ECCV*, Sep. 2016, pp. 630–645.
[12] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Aggregating cloud-free Sentinel-2 images with Google earth engine," in *Proc. Munich Remote Sens. Symp.*, 2019.
[13] X. X. Zhu and *et al.*, "So2Sat LCZ42: A Benchmark Dataset for Global Local Climate Zones Classification," *IEEE Geosci. Remote Sens. Mag.*, to be published.
[14] E. J. Hoffmann, Y. Wang, M. Werner, J. Kang, and X. X. Zhu, "Model fusion for building type classification from aerial and street view images," *Remote Sens.*, vol. 11, no. 11, p. 1259, 2019.
[15] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS Int. J. Geo-Inf*, vol. 7, no. 4, p. 129, 2018.